Patent Application of

Rajeev Sharma and

Kuntal Sengupta

for

TITLE: METHOD AND SYSTEM FOR ENHANCING THREE DIMENSIONAL FACE MODELING USING DEMOGRAPHIC CLASSIFICATION

CROSS-REFERENCE TO RELATED APPLICATIONS

This application is based on and claims priority to U.S. Provisional Application No. 60/462,809, filed April 14, 2003, which is fully incorporated herein by reference.

FEDERALLY SPONSORED RESEARCH Not Applicable

SEQUENCE LISTING OR PROGRAM Not Applicable

BACKGROUND OF THE INVENTION--FIELD OF THE INVENTION

The present invention is a system and method for human face modeling from multiple images of the face using demographics classification for an improved model fitting process.

BACKGROUND OF THE INVENTION

Three-dimensional (3D) modeling of human faces from intensity images is an important problem in the field of computer vision and graphics. Applications of such an automated system range from virtual teleconferencing to face-based biometrics. In virtual teleconferencing applications, face models of participants are used for rendering scenes at remote sites, with only the need for incremental information to be transmitted at every time instance. Traditional face recognition algorithms are primarily based on the two-dimensional (2D) cues computed from an intensity image. The 2D facial features provide strong cues for recognition. However, it cannot capture the semantics of the face completely, especially the anthropometrical measurements. Typical examples of these would be the relative length of the nose bridge and the width of the eye, the perpendicular distance of the tip of the nose from the plane passing through the eye centers and the face center, etc.

The technique discussed by Aizawa and Huang in ``Model-Based Image Coding: Advanced Video Coding Techniques for Very Low Bit-Rate Application," Proceedings IEEE, vol.83, pp.259--271, Aug. 1995 adjusts meshes to fit the images from a continuous video sequence. In a surveillance scenario, we may have only the key frames from a single, or a multiple camera system, for specific time instances. Thus the computation of optical flow between consecutive image frames, captured by each of the cameras, will not be possible.

The techniques discussed by Jebara and Pentland in ``Parameterized structure from motion for 3D adaptive feedback tracking of faces," Proceedings Computer Vision and Pattern Recognition, pp.144-150, Jun. 1997 also uses optical flow computed from consecutive frames in a video to compute the model.

Fua and Miccio in ``Animated Heads from Ordinary Images: A Least-Squares Approach," Computer Vision and Image Understanding, vol. 75, No. 3, pp. 247-259, Sep. 1999 use a stereo matching based technique for face modeling. Under multiple camera surveillance, the camera system may not be calibrated properly. This is because these cameras can be moved around, whenever required. Thus, the assumption of the knowledge of calibration parameters, especially in stereo-based techniques, breaks down.

US Pat. No. 6,556,196 describe a morphable model technique which require a frontal shots of the face. The single view based modeling approaches works well with cooperative subjects, where the entire frontal view of the face is available. Again, in vieo surveillance, it may be difficult to control the posture of the subject's face.

US Pat. No. 6,016,148 discusses a method of mapping a face image to a 3D model. The 3D model is fixed, and general. No knowledge of the demographics of the person is used, and this mapping can be erroneous, especially while using a generic model for any race or gender.

US Pat. No. 5,748,199 discusses a method of modeling three-dimensional scenes from a video, by using techniques similar to structure from motion. This technique would not be successful if continuous video feed is not provided to the system. Similar modeling technique is discussed in US Pat. No. 6,047,078. US Pat. No. 6,492,986 combines optical flow with deformable models for face modeling. As before, these techniques will not be successful when there is no continuous video stream.

US Pat. No. 5,818,959 discusses a method similar to space curving for generating three-dimensional models from images. Although these images need not be from continuous video sources, they need to be calibrated a-priori. Camera calibration is not a trivial task, especially for portable camera systems.

SUMMARY

The system first utilizes tools for face detection and facial feature detection. The face and feature detection is robust under changes in illumination condition.

Next, the system utilizes Support Vector Machine (SVM) based race and gender classifiers to determine the race and gender of the person in the images. One of the key

elements of an SVM based recognition technique is the learning phase. In the learning phase, a few thousand images for males and female faces are collected, and are used as an input for the training of the gender recognition system. Similar training procedure is followed for race classification.

For a given set of face images of the person, the race and gender is determined, and a face model, specific for that sub-class (for example, male-Caucasian is a subclass) is chosen as an approximate face model.

Next, a simple yet effective, 3D mesh adjustment technique based on some of the fundamental results in 3D computer vision was used. Fundamental results for paraperspective camera projection form the foundation of this mesh adjustment technique. Once the facial landmarks are identified across the images, the depth of an arbitrary point in the face mesh is changed continually and reprojected to all views (following paraperspective camera projection properties). The depth value for which a successful match is obtained across views is chosen. This is repeated for a dense set of points on the face.

DRAWINGS--FIGURES

FIG. 1 is an overall view of the preferred system components for the invention.

FIG. 2 is the illustration of the different important sub-components in the face modeling system.

FIG. 3 is shows the paraperspective camera model, which is approximated as the orthographic projection of points to a plane, followed by an affine transformation.

FIG. 4 is a graph for the plot of (*a4* , *ai*) over all possible images, which leads to a straight line.

FIG. 5 shows the four landmarks on the face model, the hypothetical basis plane, and the perpendicular projections of the fourth and the *i*th point.

FIG. 6 shows the plot of *a4* vs. *ai*, which is a straight line passing through (*a4* , *ai*), for a chosen value of the depth ratio.

DETAILED DESCRIPTION OF THE INVENTION

In the exemplary embodiment shown in FIG. 1, a camera, such as the Sony EVI-D30, and frame grabber, such as the Matrox Meteor II frame grabber, may be used as a means for capturing images 101. A firewire camera, such as the Pyro 1394 web cam by ADS technologies or iBOT FireWire Desktop Video Camera by OrangeMicro, or a USB

camera, such as the QuickCam Pro 3000 by Logitech, may be used as the means for capturing images 101. A plurality of such means for capturing images 101 can be used for multiple processing for multiple users 105 in the exemplary embodiment shown in FIG. 1.

Optionally, a means for displaying contents 102 in the invention can be used to render the three-dimensional face model. The means for displaying contents 102 can be any kind of conventionally known displaying device, computer monitor, or closed circuit TV. A large display screen, such as the Sony LCD projection data monitor model number KL-X9200U, may be used as the means for displaying contents 102 in the exemplary embodiments.

The processing software and application may be written in a high-level computer programming language, such as C++, and a compiler, such as Microsoft Visual C++, may be used for the compilation in the exemplary embodiment. Face detection software can be used to detect the face region 104.

In the exemplary embodiment shown in Figure 2, the system first utilizes tools 202 for face detection and facial feature detection from images 201. For the face detection and facial feature detection, any robust, reliable, and efficient detection method can be used. In U.S. Pat. No. 6,184,926 of Khosravi et al. and U.S. Pat. No. 6,404,900 of Qian et al., the authors disclosed methods for human face detection. In M.H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting Faces in Images: A Survey," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, no. 1, Jan. 2002, the authors describe various approaches for the face detection. In the exemplary embodiment, a neural

network based face detector or SVM based face detection method may be used. H. Rowley, S. Baluja, and T. Kanade, "Neural Network-Based Face Detection," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 20, no. 1, pp. 23-38, Jan. 1998, explains about the neural network based face detector in more details. E. Osuna, R. Freund, and F. Girosi, "Training Support Vector Machines: An Application to Face Detection," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 130-136, 1997 explains about the SVM based face detection approach in more details. An efficient facial feature detection is described by C.H. Lin, and J.L. Wu., "Automatic Facial Feature Extraction by Genetic Algorithms". IEEE transactions on image processing, volume 8, no. 6, pages 834-845, June 1999.

Next, the system utilizes Support Vector Machine (SVM) based race and gender classifiers, 203 and 204, respectively, to determine the race and gender of the person in the images. One of the key elements of an SVM based recognition technique is the learning phase. In the learning phase, a few thousand images for males and female faces are collected, and are used as an input for the training of the gender recognition system. Similar training procedure is followed for race classification. Examples of demographic classification for gender and ethnicity are described in detail in R. Sharma, L. Walavalkar, and M. Yeasin, "Multi-modal gender classification using support vector machines (SVMs)", U.S. Provisional Patent, 60/330,492, Oct. 16, 2001 and in R. Sharma, S. Mummareddy, and M. Yeasin, "A method and system for automatic classification of ethnicity from images", U.S. Patent, 10/747757, Dec. 29, 2003, respectively.

For a given set of face images of the person, the race and gender is determined, and a face model, specific for that sub-class (for example, male-Caucasian is a subclass) is chosen as an approximate face model by the subsystem 205 in the exemplary embodiment shown in FIG. 2.

In the exemplary embodiment shown in FIG. 2, a simple yet effective, 3D mesh adjustment technique 206 based on some of the fundamental results in 3D computer vision was used. Fundamental results for paraperspective camera projection form the foundation of this mesh adjustment technique. The paraperspective camera projection assumption works well for face modeling applications, because the depth variation on the face is not significant compared to its distance from the camera. The final face model 207 is the output of the system.

Jacobs in ``The Space Requirements of Indexing Under Perspective Projection", IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 18, no. 3, pp. 330--333, 1996, simplifies the camera projection model as an orthographic projection into a plane followed by an affine transform of these (projected) points. For a set of points (P_1, P_2 , ..., P_n) in the 3D space, a hypothetical plane passing through points P_1, P_2 and P_3 can be constructed. This is called as the basis plane, as in FIG. 3. The point P_4 is projected perpendicularly into the basis plane, and we call this projected point as p_4'. The affine coordinates of p_4' with respect to the basis (P_1, P_2 ,P_3) are (a_4, b_4). Similarly, for the ith point P_i, its projection on the basis plane is p_i' ,with affine coordinates (a_i, b_i). Parameters d_4 and d_i are the distances of points P_4 and P_i from the basis plane, respectively.

For affine coordinates (a_4, b_4) it can be shown that there is a viewpoint in which the projection of the point P_4 has those affine coordinates. The point p_{b4} lies on the basis plane with affine coordinates (a_4, b_4) for the basis (P_1, P_2, P_3). The line passing through p_{b4} and P_4 sets this viewing direction. This line meets the image plane (whose normal is parallel to the line) at a point q_4. That is, q_4 is the image of P_4. In a similar manner, P_1, P_2, P_3 are projected into q_1, q_2 and q_3, respectively on this image plane. With (q_1, q_2, q_3) as the basis, one can easily observe that q_4 has the affine coordinates (a_4, b_4), even when we subject the points on the image plane to an affine transformation (which includes translation, rotation, and scaling, to name a few).

The affine coordinates (a_i, b_i) of the projections of the remaining points (for this given view direction) are computed next as functions of (a_4, b_4). Let p_{bi} be the intersection point of the basis plane and the ray parallel to the viewing direction and passing through P_i. Let q_i be its projection on the image plane. As before, both p_{bi} and q_i have the affine coordinates (a_i, b_i) when the basis chosen are (P_1, P_2, P_3) and (q_1, q_2, q_3), respectively. Using similar triangles $P_4 p_{b4} p_{4'}$ and $P_i p_{bi} p_{i'}$ we have:

$p_{bi} - p_i = d_i (p_{b4} - p_i)/d_4$

In terms of the a affine coordinates, we express the above equation as:

$$\alpha_i = a_i + \frac{d_i}{d_4}(\alpha_4 - a_4)$$

A similar equation can be written for the b coordinate values. *The slope of the b* coordinate values is the same as that for the a affine coordinates as in Figure 4

Note that a_4, a_i , d_i and d_4 are constant over all possible images that can be generated for the given set of 3D points. Thus, for every possible image generated for (P_1, P_2 , ..., P_n) the plot of (a_4, a_i) is a straight line with a slope d_i/d_4. The straight line passes through the points (a_4, a_i) that is independent of the camera parameters, and depends solely on the 3D geometry of the points. The slope of the line is indicative of how far P_i is from the basis plane. This property will be next to estimate the structure of the human face from multiple images. Also if the equation of the affine lines are determined, then given a "target" image where we have identified the location of the projection of (P_1, P_2 , P_3 , P_4), the projection of the ith point P_i in this image can be identified by computing (a_i, b_i), using the equation of the affine lines. Repeating this for all values of i will generate the novel view of the face synthetically.

The facial feature extraction stage located the four important landmarks on the human face: the location of the eyes, nose and the mouth. Assume that the three point features (the center of the two eyes and the mouth) forms the basis, and we call them P_1, P_2 and P_3, respectively. The imaginary plane passing through these points is called the basis plane. The tip of the nose is the fourth point, P_4. These points are illustrated for the 3-D face model ias in Figure 5. Let, d_4 be the perpendicular distance of from the basis plane. Let P_i' be a point on the basis plane. Its affine coordinate values are (a_i, b_i), with (P_1, P_2 ,P_3) as the basis. If we draw a line emanating from this point and perpendicular to the basis plane, lets assume that it intersects the face model at P_i. Also, let | P_i P_i'| = d_i. Thus, given a generic 3-D CAD model of the face, we map its eyes, nose and mouth position to these features identified in the 2-D image. The task in the

face modeling stage is to estimate d_i, for the i th point on the mesh. This is repeated for all values of i.

In the kth image ($k=1,\ldots, N_f$), let the image of point P_1 be q_1^k, and so on. Consider (q_1^k ,q_2^k ,q_3^k) as the basis. From the earlier section, it is known that, for any para perspective projection of five 3-D points (P_1, P_2 , P_3, P_4, P_i) , the affine coordinates of the projection of P_4 is related to that of the projection of P_i by the equation

$$\alpha_i^k = a_i + \frac{d_i}{d_4}(\alpha_4 - a_4)$$

where (a_4^k, b_4^k) are the affine coordinates of the projection of P_4 in the kth view, and so on.

The right hand side of the equation is only a function of the unknown parameter $s_i = d_i/d_4$, which we formally call the *depth ratio*. Here, a_4 is known and is a race and gender dependent constant. The a_i^k component can be estimated similarly as a function of s_i. Next, we compute ($x_i^k(s_i)$, $y_i^k(s_i)$), the image coordinate values in the kth frame. The average sum of the squared difference measure of the intensity as a function of s_i, computed over every image pair chosen, is defined as follows.

$$SSD(s_i) = \frac{2}{N_f(N_f - 1)} \sum_{k=1}^{N_f - 1} \sum_{l=k+1}^{N_f} DIFF(win(k, x_i^k, y_i^k, w), win(l, x_i^k, y_i^k, w))$$

Here $win(k, x_i^k, y_i^k, w)$ is a window of size wxw selected in the kth image around the point (x_i^k, y_i^k). Also, *DIFF(.)* is the sum of the squared difference computed for the window pair.

The estimated value s_i is the one for which SSD(s_i) is minimum. Theoretically, one has to search from $[-\infty, \infty]$. In the system the search is constrained as follows. After the 3D model is fitted to the face for the ith point, if the depth ratio according to this generic model is s_i^0, then we search in the neighborhood of this value. The search can typically be constrained in the neighborhood of s_i^0 .

The depth ratio estimation process can be interpreted graphically as in Figure 6. The straight line corresponding to the a_4 vs. a_i plot always passes through the point (a_4, a_i), and the slope of the line is the unknown parameter s_i we seek to estimate. The slope is varied over a range of values. For a particular setting of the slope value, a_i^k for a given a_4^k is generated. The depth ratio estimation process is repeated for a dense set of points on the basis plane with affine coordinates (a_i, b_i), following the steps discussed earlier. The next issue is to obtain the Euclidean coordinate values of the ith point starting from the parameters (a_i, b_i, s_i), which we refer to as the affine structure of P_i.

With the knowledge of the Euclidean geometry of certain reference points, such as distances and angle values, it is possible to estimate Euclidean structure of all the points on the mesh by minimizing a penalty function. For face modeling application, the Euclidean coordinate values of the template model's eyes, nose and mouth position are used, from which the Euclidean structure of the subject's face is generated. Next, using the texture from one of the input images, the face can be rendered for different pitch and yaw values (i.e., rotation in x- and y- axis).

The final system allows the derivation of anthropometric measurements from facial photographs taken in uncontrolled or poorly controlled conditions of resolution, pose angle, and illumination.